Paper

Supervised Kernel Principal Component Analysis by Most Expressive Feature Reordering

Krzysztof Ślot, Krzysztof Adamiak, Piotr Duch, and Dominik Żurek

Institute of Applied Computer Science, Lodz University of Technology, Lodz, Poland

Abstract—The presented paper is concerned with feature space derivation through feature selection. The selection is performed on results of kernel Principal Component Analysis (kPCA) of input data samples. Several criteria that drive feature selection process are introduced and their performance is assessed and compared against the reference approach, which is a combination of kPCA and most expressive feature reordering based on the Fisher linear discriminant criterion. It has been shown that some of the proposed modifications result in generating feature spaces with noticeably better (at the level of approximately 4%) class discrimination properties.

Keywords—feature selection, kernel methods, pattern classification.

1. Introduction

Kernel methods [1], [2] enable derivation of highly discriminative feature spaces by linearizing class separation problems in implicitly-exploited, very high-dimensional spaces. Adoption of the optimal feature space is a key issue in pattern recognition, as majority of real-world pattern recognition problems are typically highly nonlinear, and many diverse nonlinear approaches for handling this issue have been proposed so far such as locally linear embeding [3], Laplacian Eigenmaps [4] or Isomaps [5].

Several important concepts in the field of kernel-based feature space derivation have been formulated so far. A basic scheme for derivation of a nonlinear feature space with kernel methods is an extension to the classical Principal Component Analysis (PCA) method. This scheme, named kernel-PCA and proposed by Scholkopf *et al.* in [6], produces a feature space from a subset of most-expressive features (MEF) determined for projections of original samples onto a nonlinear, high-dimensional Hilbert space. A rationale behind that scheme is the same as in case of a regular PCA: large data scatter is likely to be produced by separable clusters, possibly belonging to different classes. As MEF-based feature space derivation has obvious limitations, feature selection as well as feature extraction schemes have been developed to improve classification performance of kernel methods. In case of the latter direction, the two most notable methods proposed so far are: kernel Fisher Discriminant Analysis (kFDA), formulated in [7], and Supervised Principal Component Analysis (SKPCA), proposed by Barsham *et al*. in [8]. The kFDA generalizes classical linear discriminant analysis for kernel-induced spaces where it determines a direction of maximum linear class separability. On the other hand, SKPCA produces an ordered set of the most discriminative directions, defined as the ones that maximize Hilbert-Schmidt norm of cross co-variance matrix, which describes relations between projected samples and their class labels. Both approaches proved extremely successful, however, there exist aspects that could potentially challenge their high performance. The main potential issue that exists in case of kFDA is the resultant one-dimensional output space. This problem can become more serious than in case of Support Vector Machine (SVM) classification [9], as no maximum margin criterion is involved in search for the most discriminative direction, generated by kFDA. SKPCA bypasses the aforementioned issue, however, there exist no clear guidance on selection of quantitative class labels and their kernels, which are important components of the method.

The main reasons for considering feature selection performed on kPCA results as a promising feature space derivation strategy are the following. The first advantageous property of such an approach, which does not hold for SVM classification or kernel Fisher Discriminant Analysis, is a presence of a broad pool of mutually orthogonal candidates that could build a multidimensional discriminative space, which would host projections of class samples. Moreover, as classification problems tend to get linear in kernel-induced feature spaces, even linear feature selection criteria applied in these spaces could provide good assessment of class separation. Finally, one needs to keep in mind that feature selection is performed on results of kPCA analysis, which means that each feature of a target space is some nonlinear combination of all original features, so complete information on the problem embedded in input data is used, as opposed to the case of conventional feature selection, performed directly in input space, where information from dropped features is inevitably lost.

The presented research is aimed at exploring methods for discriminative feature space derivation, which depart from results of kernel-PCA of input datasets. A strategy adopted for the task realization is feature selection, where features are eigenvectors of projected sample distributions (through kernels) that exist in high-dimensional spaces, henceforth referred to as $\mathscr{H}$ space. Feature selection in $\mathscr{H}$ space,

i.e. selection based on kernel-PCA results, have already been addressed in several publications. For example, unsupervised approach to feature selection in $\mathscr{H}$ space has been proposed in [10]. Supervised selection of features in kPCA-produced space have been considered in [11], [12].

The main contribution of the paper is exploration of a set of feature selection criteria and verification of performance of corresponding, derived feature spaces. The proposed criteria are in general nonlinear, and they are applied to nonlinear projections of original samples onto k-PCA derived directions.

The problem of multiple-category classification has been addressed in the presented research. Several different ways of class separation scoring were considered to evaluate candidate $\mathscr{H}$ space directions. The first criterion for recruiting target space features seeks for directions that maximize balanced class separation, assessed over all classes. The second one favours directions that provide the maximum pairwise class separation. Both criteria are also subject to modifications that emphasize class distribution divergence from symmetry and Gaussianity. The presented methods are confronted with kernel-PCA based classification and k-NN (k-nearest neighbor) classification, performed in the original feature space.

The paper is organized in the following way. A background for the presented research, including a brief review of kernel PCA is outlined in Section 2. The proposed feature selection strategies and criteria are described in Section 3. Section 4 presents assessment methodology used for the proposed concepts and provides results of methods experimental evaluation.

2. Related Work

The proposed methods are based on the theory of kernel PCA and on classical theory of Fisher linear discrimination. As Fisher's linear discrimination theory is one of the fundamental and well-known concepts in pattern recognition (see e.g. [13]–[16]), only kernel principal analysis has been outlined in the remaining part of this Section.

2.1. Kernel Principal Component Analysis

Kernel Principal Component Analysis attempts to find directions of the maximum scatter of data projected onto some high-dimensional (possibly, infinitely-dimensional) feature space. Denoting a set of data samples, defined in an original low-dimensional space $\mathscr{L}$ by $\{\mathbf{x}\}$ and introducing a function $\Phi(.)$ that transforms these samples onto another space, of higher dimension ($\mathscr{H}$), the projections $\mathbf{X}_i$ of original samples $\mathbf{x}_i$ are given by:

$$\mathscr{L} \rightarrow \mathscr{H}: \qquad \mathbf{X}_i = \Phi(\mathbf{x}_i). \tag{1}$$

The PCA problem for the samples $\mathbf{X}_i$ arranged in the matrix $\mathbf{X}$ can be stated as:

$$(\mathbf{X}-\mathbf{M})(\mathbf{X}-\mathbf{M})^T\mathbf{V} = n\Lambda\mathbf{V}, \tag{2}$$

where Λ and $\mathbf{V}$ are eigenvalue and eigenvector matrices respectively, n is the number of samples and $\mathbf{M}$ denotes a matrix of projection mean vectors μ, i.e.:

$$\mathbf{M} = [\mu, \mu \ldots]. \tag{3}$$

By premultiplying both sides of the Eq. (2) by the term $(\mathbf{X}-\mathbf{M})^T$ and observing that each eigenvector $\mathbf{V}_k$ exists in a space spanned by original data projections (is a linear combination of samples $\mathbf{X}_i$), i.e.:

$$\mathbf{V}_k = \sum_{i=0}^{n-1} \alpha_k^i(\mathbf{X}_i - \mu) \rightarrow \mathbf{V}_k = (\mathbf{X}-\mu)\alpha_k, \tag{4}$$

where α_k^i are weights normalized so that the vector $\alpha_k = [\alpha_k^0, \alpha_k^1, \ldots]$ is of the unit length, the Eq. (2) can be restated in the form:

$$\mathbf{G}\mathbf{A} = n\Lambda\mathbf{A}, \tag{5}$$

where $\mathbf{G}$ is the Gram matrix computed for projected samples:

$$\mathbf{G} = (\mathbf{X}-\mathbf{M})^T(\mathbf{X}-\mathbf{M}), \tag{6}$$

and $\mathbf{A}$ is a matrix hosting vectors α_k, i.e. $\mathbf{A} = [\alpha_\mathbf{0}, \alpha_\mathbf{1}, \ldots]$, which can be seen as a matrix of parametric representations of eigenvectors of the system given by Eq. (2).

As dot products of vectors in high-dimensional space $\mathscr{H}$ are involved in derivation of Eq. (5), one can apply a kernel function $k(.)$ (providing that it exists) and perform all the computations using data from the original space:

$$< \Phi(\mathbf{x}_i) - \mu, \Phi(\mathbf{x}_j) - \mu > = k(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

Centering of high-dimensional samples around the mean, which is crucial for searching for most expressive features in $\mathscr{H}$ space, can be done by an appropriate modification of the $\mathbf{G}$, yielding $\mathbf{G}_c$. This leads to the final formulation of the kPCA:

$$\mathbf{G}_c\mathbf{A} = n\Lambda\mathbf{A}. \tag{8}$$

Projections of unknown samples $\mathbf{x}_p$ onto eigenvectors derived for the $\mathscr{H}$ space, can be also computed using kernels, as they involve sums of dot products:

$$y_p^k = < \Phi(\mathbf{x}_p), \mathbf{V}_k > = \sum_{i=0}^{n-1} \alpha_k^i \Phi(\mathbf{x}_p)^T\Phi(\mathbf{x}_i) = \sum_{i=0}^{n-1} \alpha_k^i k(\mathbf{x}_p, \mathbf{x}_i). \tag{9}$$

The most frequently used kernel functions, which are also considered in the presented research, are Gaussians, polynomials and extended polynomials. Gaussian kernel transforms samples into infinitely-dimensional space $\mathscr{H}$. It involves one parameter σ, which needs to be appropriately chosen [17], and it is defined as:

$$k_G(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)}{\sigma^2}}. \tag{10}$$

Polynomial kernels are defined as:

$$k_p(\mathbf{x}_i, \mathbf{x}_j) = < \mathbf{x}_i, \mathbf{x}_j >^m, \tag{11}$$

and

$$k_x(\mathbf{x}_i, \mathbf{x}_j) = (< \mathbf{x}_i, \mathbf{x}_j > +1)^m, \tag{12}$$

where m is the polynomial order and the symbol $< .,. >$ stands for a dot product.

3. Feature Selection Criteria

Kernel PCA finds a set of orthogonal vectors that maximize scatter of original sample projections in $\mathscr{H}$ space. Since unlabeled samples are used in data analysis, most expressive directions might not correlate with class-separability (as it is the case for the conventional PCA). Sample results of application of kPCA to artificially generated, two-class data set have been presented in Fig. 1. Projections of original samples onto the first two most expressive features, shown in Fig. 2, clearly show that kPCA cannot provide good data representation for class discrimination. Therefore, selection of features produced by kPCA, aimed at derivation of discriminative spaces for data classification, has been considered, and various feature selection criteria have been proposed and examined in what follows.

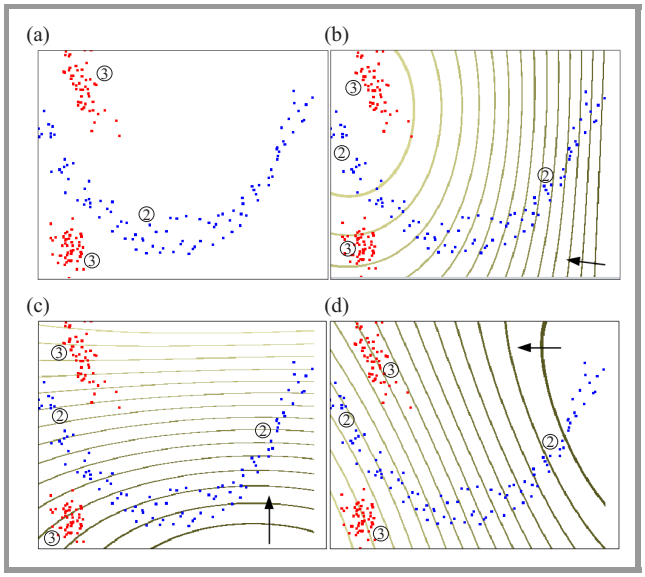

Fig. 1. Sample two-dimensional distributions of two classes, denoted by "2" and "3" (a), superimposed with isoclines that correspond to kernelized dot products of domain points with the first eigenvector (b), with the second one (c) and with the third one (d). The arrows indicate increasing values of a dot product.

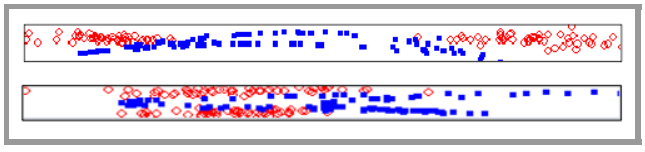

Fig. 2. Projections of data samples onto the first (top) and the second (bottom) most expressive feature. To increase clarity, samples are vertically dispersed inside a band of non-zero height.

The adopted feature selection methodology is built on the simplest setup. The various criteria for individual KPCA-produced features are formulated and the best performing ones are chosen to the resulting subspace. The applied criteria are based on the principle underlying Fisher Linear Discriminant Analysis [14], which is maximization of between-class to within-class scatters. The authors investigate a set of different particular definitions of these quantities. As they consider a multiple-category classification, the basic formulation for a feature selection criterion has the form:

$$F_1^{\xi} = \frac{det\left(\sum_{i=0}^{n-1}(\mathbf{M}_i^{\xi} - \mathbf{M}^{\xi})(\mathbf{M}_i^{\xi} - \mathbf{M}^{\xi})^T\right)}{\sum_{i=0}^{n-1} det(\mathbf{C}_i^{\xi})}, \qquad (13)$$

where n is the number of classes, $det(\mathbf{C}_i^{\xi})$ denotes a determinant of co-variance matrix for projections of i-th class samples onto some ξ-th subspace of the $\mathscr{H}$ space, $\mathbf{M}_i^{\xi}$ denotes a mean vector for projections of i-th class samples and $\mathbf{M}^{\xi}$ is the mean of class means $\mathbf{M}_i^{\xi}$. For one-dimensional case (when a single feature is to be evaluated) the criterion (13) can be expressed in a form that employs simplified measures of within-class and between-class scatters:

$$F_1^k = \frac{\sum_{i=0}^{n-1} |\mu_i^k - \mu^k|}{\sum_{i=0}^{n-1} \sigma_i^k}, \qquad (14)$$

where σ_i^k is a standard deviation of projections of i-th class samples onto k-th feature, μ_i^{ξ} is a mean of i-th class sample projections onto k-th feature and μ^k is the mean of means.

Given the feature scores produced by Eq. (14), the first criterion for feature space derivation, resulting in D-dimensional most-discriminative feature set $\mathbf{F}_1$, can be expressed as:

$$\mathbf{F}_1 = \{F_1^{\alpha_0} \ldots F_1^{\alpha_{D-1}}\} : \alpha_d = arg \max_{k \neq \alpha_0 \ldots \alpha_{d-1}} (F_1^k). \qquad (15)$$

Results of most expressive feature reordering, based on criteria (13) and (14), are summarized in Figs. 3 and 4, where three-class distributions were processed according to two different scenarios. In the first case, original samples were subject to kPCA analysis, where Gaussian kernel (10) was applied (a value of $\sigma = 2$ was used), and three most expressive features were selected as a subspace for projected data classification. As it can be seen from Fig. 4, distributions of projections of considered class samples remain nonlinearly bounded (with concave bounding surfaces). A very different situation is presented if feature selection is used. This time features with indices 0, 9 and 5 were selected (increasing feature index corresponds to a decreasing data scatter in the corresponding direction). As it can be seen from isoclines drawn in Fig. 3 the eigenvectors segment the original two-dimensional domain in much more complex way, which is beneficial from the point of view of data separation. This can be seen in Fig. 4, where three dimensional feature space provides a very simple structure to class distributions – they become linearly separable.

The criteria (13) and (14) seek for a simultaneous assessment of distribution separability for all classes, using an ambiguous score for between-class scatter (numerator).

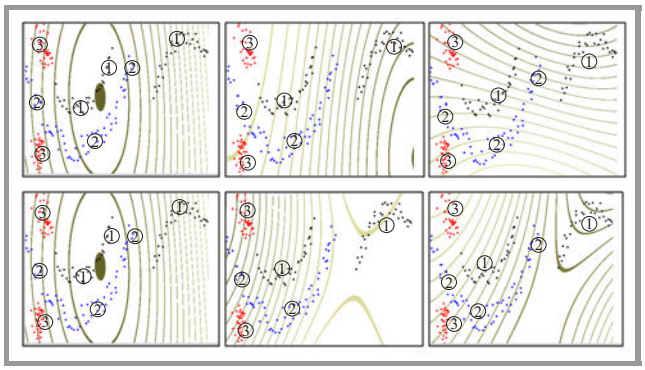

Fig. 3. Iscoclines defined by a set of constant values of kernelized dot-products between KPCA eigenvectors and data 2D domain points. Results of space labeling using the three most expressive vectors (top row), derived using KPCA analysis of input data and space labeling with most discriminative features, according to the criterion (14), with indices: 0.9 and 5, respectively (bottom row). Points of the three classes are shown in black (1), blue (2) and red (3).

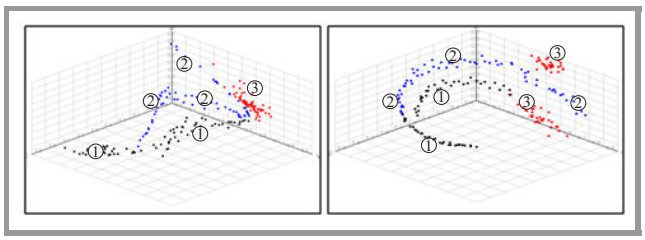

Fig. 4. Representation of original samples in three-dimensional feature spaces, defined by three most discriminant (in a sense of the criterion (14) eigenvectors from a set of KPCA results (left) and the space defined by the three most expressive feature vectors (right).

The score favors evenly spaced Gaussian class distributions, which is rarely the case in practice. Therefore, the authors propose to introduce scoring of pairwise class separation only, and to build a target feature space from a collection of directions that provide best separations for all pairs of classes. This approach might potentially lead to a large feature space cardinality if large number of classes are considered, however, it has been found that individual directions typically provide the best separation for several class pairs. The proposed feature selection criterion produces the most-discriminative feature set $\mathbf{F}_2$:

$$\mathbf{F}_2 = \{F_2^{\alpha_0} \ldots F_2^{\alpha_k} \ldots F_2^{\alpha_{D-1}}\}, \tag{16}$$

where its elements $F_2^{\alpha_k}$ provide maximization of a separation score between classes p and q, assessed using the score:

$$F_2^{\alpha_k} = \frac{|\mu_p^{\alpha_k} - \mu_q^{\alpha_k}|}{\sigma_p^{\alpha_k} + \sigma_q^{\alpha_k}}. \tag{17}$$

The second modification that has been introduced is concerned with tuning feature scoring criteria, so that kPCA-produced features where projected samples have distributions that are actually close to Gaussian, become the preferred ones. The assumption of distribution Gaussianity is at the core of all of the presented feature selection criteria, however it is not justified in any way. As a result, class samples typically remain mixed even though their Gaussian models, expressed using means and standard deviations, suggest decent class separability. To assess actual properties of projected sample distributions two different scores for within-class scatter assessment are introduced. To penalize heavily asymmetric distributions (with long tails that can mix with samples from apparently distant classes) the distribution skewness s is included, i.e., the third central moment, into denominators of class separation scores, so that the corresponding criteria (14) and (17) assume the following forms:

$$F_{1S}^k = \frac{\sum_{i=0}^{n-1} |\mu_i^\xi - \mu^\xi|}{\sum_{i=0}^{n-1} \sigma_i^k (1 + |s_i^k|)}, \tag{18}$$

and

$$F_{2S}^{\alpha_k} = \frac{|\mu_p^k - \mu_q^k|}{\sigma_p^k (1 + |s_p^k|) + \sigma_q^k (1 + |s_q^k|)}, \tag{19}$$

where s_i^k denotes skewness of i-th class samples projection onto some k-th eigenvector. Observe that the proposed modification is penalizing asymmetric distributions, by reducing the corresponding scores. Similarly, to prefer Gaussianity of distributions, kurtosis κ is included in an analogic manner into these criteria, yielding:

$$F_{1K}^k = \frac{\sum_{i=0}^{n-1} |\mu_i^\xi - \mu^\xi|}{\sum_{i=0}^{n-1} \sigma_i^k (1 + |\kappa_i^k|)} \tag{20}$$

and

$$F_{2S}^{\alpha_k} = \frac{|\mu_p^k - \mu_q^k|}{\sigma_p^k (1 + |s_p^k|) + \sigma_q^k (1 + |s_q^k|)}. \tag{21}$$

As a final remark, the authors would like to emphasize that all sample separation criteria introduced in Section 3 also hold in original feature spaces, without a necessity to perform nonlinear, kernel-based transformations.

4. Experimental Evaluation of the Considered Strategies

Experimental setup used for verification of the proposed concepts was the following. Four-category classification problem was considered with artificially generated samples, defined in 25-dimensional space, of which only 3-dimensions provided structured class distributions (see Fig. 5). In this 3D subspace, distribution of three of the considered classes (shown in black, red and blue and marked as 1, 2, 3, respectively) were bimodal. The fourth class distribution (shown in green, marked by 4) fills in a concave region in space, which encloses one of the modes of the red as well as of the blue class. Also the other

modes of red and blue classes occupy concave regions. For the remaining twenty two dimensions, sample coordinates were generated randomly (with either uniform or binomial distributions), thus making these directions useless from the point of view of class discrimination. One thousand-element set of samples was generated, including even number of samples (i.e. 250) per class.

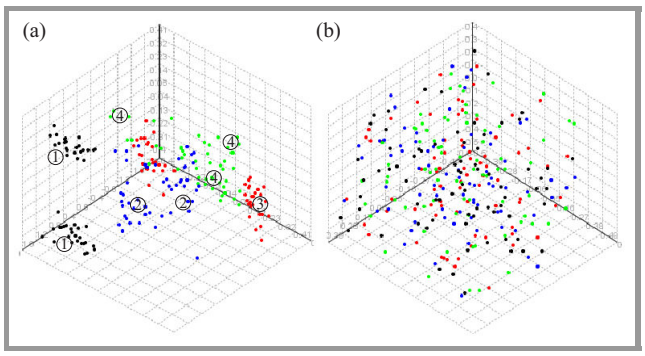

Fig. 5. Two projections of a generated, 25-dimensional distribution of input samples, onto three-dimensional subspaces: (a) the only subspace with structured data distribution, (b) a 3D subspace selected randomly from the remaining twenty-two dimensions.

An objective of the experiments was to evaluate discriminative properties offered by different feature spaces. An outcome of k-NN classification, performed in the 10-fold cross-validation scheme (in each run, training and test samples were mutually exclusive) was considered for feature space scoring. A particular choice of the k-NN strategy was made because the considered feature space derivation methodology is weakly correlated with k-NN classification principles. To provide the reference results for comparison against performance of the considered kernel-based strategies, performance of k-NN classification applied to raw input data was also evaluated.
The following experimental setup was adopted. Each of the feature selection criteria, followed by k-NN classification was performed on the same set of artificially generated data. For each procedure, a set of alternative parameters was used, including:

- a type of the kernel (the kernels given by Eqs. (10), (11) and (12) were considered) and its parameter values (orders, for polynomial kernels and σ for the Gaussian kernel),
- target feature space cardinality (denoted henceforth by D),
- classification method parameter k.

Sample output data distributions in target 3D feature spaces, derived using three different methods: basic kPCA and two spaces obtained by application of feature selection procedure, involving the criteria F_1 (14) and F_2 (17), have been shown in Fig. 6. Although samples do not form clear clusters and no substantial differences can be observed among the plots, classification performance in these spaces is quite different, starting from 61.5% for the first space, through 68% for the second one, to 74.5% for the third one.

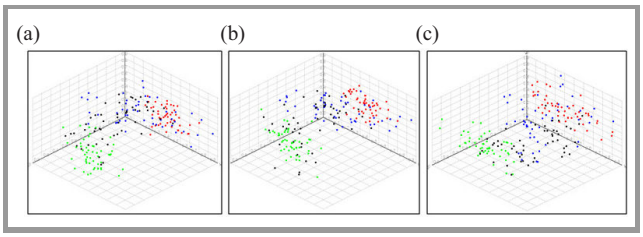

Fig. 6. Distribution of samples projected onto axes of a target feature space, derived using: (a) KPCA method, (b) feature selection driven by the score F_1, (c) feature selection driven by F_2.

Performance of the reference method – data classification in input 25-dimensional space using k-NN method – equals 64.5%. In all cases, a value of $k = 7$ was used. The presented results appear to be characteristic for the considered methods.
An extensive summary of experiments aimed at evaluating class discrimination performance of different feature spaces, is provided in the following tables and figures, where the following notation has been adopted for the considered feature spaces:

- kPCA – denotes a space composed of a set of most expressive features, i.e. the leading eigenvectors derived using kernel-PCA),
- F_1 – denotes a space composed of the most discriminative vectors derived using the feature selection criterion (14),
- F_2, F_{2S}, F_{2K} – denote spaces composed of most discriminative vectors derived using the feature selection criterion (17) and its modifications involving skewness (19) and kurtosis (21), respectively,
- RAW – denotes the original feature space.

The first group of experiments was concerned with comparison of performance of data classification in spaces derived using methods F_1 and F_2, in confrontation with a space derived using kPCA and data classification by means of k-NN in the original space (RAW). A target space dimensionality of $D = 3$ was assumed and Gaussian kernel with $\sigma = 2$ was chosen (the choices were not optimized in any way). Results, presented in Table 1, confirm the expectation that feature space built from most expressive features performs poorly in class separation. One can observe that it is also outperformed by k-NN data classification performed on raw data (classification performance increases with cardinality k of a winner set). Feature selection, as expected, performs the best, however to justify computational overhead necessary for feature space derivation, tuning of kernel parameters was necessary.
The three kernels presented earlier: Gaussian (10), polynomial (11) and extended polynomial (12) were tested during

Table 1
Classification performance in various target feature spaces for fixed target feature space dimension $D = 3$ and without optimization of procedure parameters

k	RAW	kPCA	F_1	F_2
1	66.2	56.5	72.2	72.5
3	66.2	59	**74.2**	72.2
5	64.2	62.2	72.7	73.2
7	71.2	**64.5**	73.5	**74**
11	**72.1**	62.2	72	72.7

the following experiments. For the Gaussian kernel, parameter tuning was reduced to choice of the parameter σ, which determines a range of training set points influence. In case of polynomial kernels, the tuning concerned polynomial order (parameter m of equations (11) and (12)). Three different values for m were tested throughout experiments: $m = 2$, 3 and 6. One needs to note, that due to high dimension of the original data vectors ($d = 25$), even for the considered low polynomial orders, a resulting feature space, where classification gets performed, has very high dimensionality. As it was shown in [9], cardinality of the $\mathscr{H}$ space, in case the polynomial (11) is considered, is related to a polynomial order m and to an original input vector dimension d via the formula:

$$D = \binom{d+m-1}{m} = \frac{(d+m-1)\cdot\ldots\cdot d}{m!}, \qquad (22)$$

which, for the considered parameters, gives $\mathscr{H}$ space dimensions: $D = 325$ for $m = 2$, $D = 2925$ for $m = 3$ and $D = 593775$ for $m = 6$.

Results of the experiments are shown in Fig. 7. One can see that Gaussian kernels outperform the polynomial ones for appropriately chosen values of the parameter σ. In case of polynomial kernels, one can notice that increase in complexity of class-separation boundaries, caused by increasing

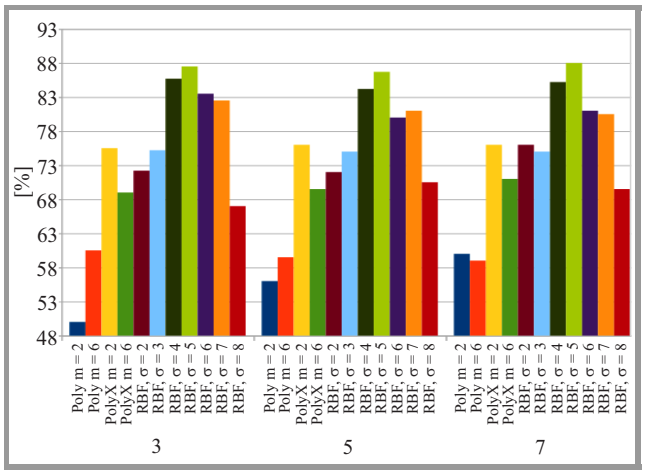

Fig. 7. Classification performance for different kernels as a function of varying k-NN classification parameter k (target space of dimension $D = 3$ is used, RBF denotes the Gaussian kernel).

the polynomial order, impairs classification performance. This is clearly a result of poorer generalization properties of higher-order curves that tend to overfit training data samples.

To compare performance of different scoring methods (involving basic class separation measures: (17) and (14), involving additionally skewness (18) and (19) and kurtosis (20) and (21)), further experiments were performed only for Gaussian kernels with a value of σ set to five. Two different cardinalities of target feature spaces: $D = 3$ and $D = 4$, were considered. The former choice was motivated by actual dimensionality of the original problem (structured, separable data exist in three-dimensional space), and the adopted feature space derivation methodology was expected to infer this dimension. The latter dimension was used for reference to see, whether classification performance in overly-dimensional space is indeed lower.

Results of the experiments performed for feature selection based on pairwise separation assessment (F_2 strategy) are shown in Fig. 8 (results for the strategy F_1 were similar). As

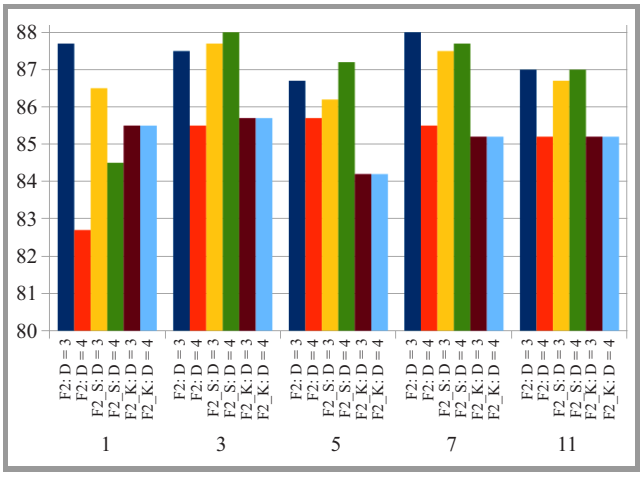

Fig. 8. Performance comparison for different variants of feature selection by pairwise separation maximization with the three considered scoring criteria: basic (17), involving skewness (19) and involving kurtosis (21), for two different dimensions of target feature spaces ($D = 3$ and $D = 4$) as a function of the k-NN classification method parameter k.

it can be seen, classification performance of the considered feature scoring methods varies, and no clear conclusion can be drawn from the resulting plots. One needs to bear in mind, that original class distributions generated in a subspace shown on the left of Fig. 5, i.e. for a 3D subspace, where distributions are structured, are uniform. The results from Fig. 8 indicate that statistical properties seem to be typically propagated to high-dimensional spaces, as a basic feature scoring criterion (17) usually performs better than the criterion that involves skewness. On the other hand, superiority of the reconstructed three-dimensional space over its four-dimensional counterpart is evident, which proves the expectations.

The final group of experiments was aimed at comparative evaluation of the two considered feature selection strate-

gies: selection by multiple-class separation assessment (F_1 strategy) and selection by pairwise-class separation assessment (F_2). Three-dimensional target feature space was assumed and projections onto $\mathscr{H}$ space was done by Gaussian kernel with $\sigma = 5$. As skewness and kurtosis did not prove to have an advantage as modifiers in feature scoring, only basic forms of selection criteria (14) and (17) were used. Experiment results are presented in Table 2.

Table 2
Comparison of classification performance in feature spaces derived using F_1 and F_2 selection criteria for Gaussian kernel and fixed dimension $D = 3$

.

k	F_1	F_2
3	83.3	**87.8**
5	84.6	**87.2**
7	83.5	**88**

As it can be seen, feature selection driven by pairwise class separation assessment performs better than feature selection driven by multiple-class separation criterion. A difference is relatively small, yet consistent. Both methods outperform k-NN classification of original samples, which tops at 72.1% for 11-NN classification (see Table 1). Also, both methods outperform classification in a feature space derived by class separation assessment in two-, three- and four-dimensional subspaces, defined by the criterion (15) (performance of this soring method was even below the one of the kPCA method). Better performance of F_2 over F_1 scheme may result from collective cooperation of features that provide good class-wise separation in a multidimensional space, resulting in correct tackling of the multi-class problems.

5. Conclusions

Different methods for feature space derivation by selection of eigenvectors produced by kernel-PCA have been examined in the presented paper. It has been shown that one can improve classification performance by introducing appropriate modifications to the feature selection procedure. The modifications that contribute to higher classification rates include reformulation of a feature selection criterion, which focuses on evaluation of pairwise class separation. Some of the proposed modifications, such as inclusion of sample distribution skewness and kurtosis int intra-class scatter scoring criteria, proved to be inconclusive.

One needs to keep in mind that experiments were performed on an artificially-generated datasets with some particular properties. Although the proposed class distributions reflect main properties of hard, real data sets, such as multi-modality, nonlinear class boundaries (including concave ones) and a significant degree of randomness, much more experiments have to be made to confirm the observed properties of the considered methods. Also, the proposed methods need to be evaluated on real datasets. Despite this, the authors believe that the results obtained provide an interesting alternative to the commonly used feature selection approaches in kernel-induced feature spaces.

Acknowledgements

The presented research has been supported by the Polish National Science Center under the research grant: 2012/05/B/ST6/03647.

References

[1] T. Hofmann, B. Scholkopf, and A. Smola, "Kernel methods in machine learning", *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[2] B. Scholkopf and A. Smola, *Learning with Kernels*. Cambridge: MIT Press, MA, 2002.

[3] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, vol. 290, pp. 2323–2326, 2000.

[4] M. Belkin and P. Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation', *Neural Comput.*, vol. 15, no. 6. pp. 1373–1396, 2003

[5] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for non-linear dimensionality reduction", *Science*, vol. 290, pp. 2319–2323, 2000.

[6] B. Scholkopf and A. Smola, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Comput.*, vol. 10, pp. 1299–1319, 1998

[7] S. Mika, G. Ratsch, J. Weston, B. Scholkoph, and K. Mullers, "Fisher discriminant analysis with kernels", in *Proc. IEEE Conf. of Neural Netw. for Sig. Process.*, Madison, WI, USA, 1999, pp. 41–48.

[8] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds", *Pattern Recogn.*, vol. 44, pp. 1357–1371, 2011.

[9] C. Burges, "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowl. Discovery*, vol. 2, pp. 121–167, 1998.

[10] M. Wang, S. Fei, and M. I. Jordan, "Unsupervised kernel dimension reduction", in *Proc. Conf. Adv. in Neural Inform. Process. Systems NIPS 2010*, Vancouver, BC, Canada, 21010, vol. 23, pp. 2379–2387.

[11] Le Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization", *J. Machine Learn. Res.*, vol. 13, pp. 1393–1434, 2012.

[12] G. Baudat and F. Anouar, "Feature vector selection and projection using kernels", *Neurocomputing*, vol. 55, pp. 21–38, 2003.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd edit. Wiley, 2000.

[14] R. A. Fisher, "The use of multiple measures in taxonomic problems", *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[15] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images,", *J. Optic. Society of America A*, vol. 14, pp. 1724–1733, 1997

[16] W. Skarbek, K. Kucharski, and M. Bober, "Dual LDA for face recognition", *Fundamenta Informaticae XXI*, vol. 1, pp. 1–33, 2001.

[17] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines", *Machine Learn.*, vol. 46, pp. 131–159, 2002.

Krzysztof Ślot is a professor at the Institute of Applied Computer Science, Lodz University of Technology Poland. His research interests include pattern recognition, speech processing, computational intelligence and human-computer interfacing.

E-mail: krzysztof.slot@p.lodz.pl
Institute of Applied Computer Science
Lodz University of Technology
Stefanowskiego st 18/22
90-924 Lodz, Poland

Piotr Duch is an adjunct professor at the Institute of Applied Computer Science, Lodz University of Technology Poland. His research interests include fractional calculus of non-integer order, pattern recognition and speech processing.

E-mail: pduch@kis.p.lodz.pl
Institute of Applied Computer Science
Lodz University of Technology
Stefanowskiego st 18/22
90-924 Lodz, Poland

Krzysztof Adamiak is a Ph.D. student at the Institute of Applied Computer Science, Lodz University of Technology Poland and also a software developer at Comarch Technologies. His research interests include image processing data classification and virtual and augmented reality system design.

E-mail: krzysztof.adam.adamiak@gmail.com
Institute of Applied Computer Science
Lodz University of Technology
Stefanowskiego st 18/22
90-924 Lodz, Poland

Dominik Żurek is a Ph.D. student at the Institute of Applied Computer Science, Lodz University of Technology Poland. His research interests include image processing data classification, human-computer interaction and artificial intelligence.

E-mail: dom.zurek@gmail.com
Institute of Applied Computer Science
Lodz University of Technology
Stefanowskiego st 18/22
90-924 Lodz, Poland